Transformer-based Direct Speech-to-speech Translation with Transcoder

Takatomo Kano¹, Sakriani Sakti¹,², and Satoshi Nakamura¹,²

1. Nara Institute of Science and Technology, Japan
2. RIKEN, Center for Advanced Intelligence Project AIP, Japan
Introduction

The traditional approach solves the speech-to-speech translation step by step. However, there are many limitations.

- Cascade of ASR, MT, and TTS using text.
- All of which is independently trained and tuned.
Existing work: Multi-task Speech-to-Speech Translation [Ye et al., 2019]

Multi-task model

Training step

- The model has three decoders but uses only one decoder for translation.
- The model uses one attention module for alignment source and target speech.
- ASR and MT decoders do not work in the inference step.
- There is no information sharing from ASR and MT.
- All modules use RNN.

Inference step

- The model uses one attention module for alignment source and target speech.
- ASR and MT decoders do not work in the inference step.
- There is no information sharing from ASR and MT.
- All modules use RNN.
Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder

The model utilizes all trained modules to translate.

- ASR and MT decoders provide attention alignment necessary in the next block.
- The model uses a combination of 3 attentions for alignment source to target speech.
- Transcoder converts different hidden states in speech translation.
- The model converts ASR and MT hidden states into MT and TTS hidden states.
- We revamp the RNN-based speech-to-speech translation model by the Transformer.
Experiment Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Syntactic similar</th>
<th>Syntactic distant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En to ES</td>
<td>Ja to Ko</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Baseline: Cascade (RNN)</td>
<td>38.9</td>
<td>47.7</td>
</tr>
<tr>
<td>Baseline: Cascade (Transformer)</td>
<td>41.3</td>
<td>52.1</td>
</tr>
<tr>
<td>Multi-task (RNN)</td>
<td>38.8</td>
<td>48.2</td>
</tr>
<tr>
<td>Multi-task (Transformer)</td>
<td>43.1</td>
<td>58.8</td>
</tr>
<tr>
<td>Transcoder (Transformer)</td>
<td>44.0</td>
<td>59.3</td>
</tr>
</tbody>
</table>

✓ Our proposed approach outperforms Multi-task speech translation.
✓ The proposed method with Transformer further improves performances.
End of my highlight talk.
Limitations in Existing Approach

Many existing works only perform on syntactically similar language pairs.

• Syntactically similar language pairs have
  • Similar word ordering (SVO-SVO or SOV-SOV)
  • A lot of one-to-one monotonic alignments.
    • English-French, English-Spanish, Japanese-Korean

• Syntactically distant language pairs have
  • Different word ordering (SVO-SOV)
  • A lot of many-to-many alignments.
    • English-Japanese
  • Difficult to translate

Direct End-to-End Speech Translation for Distant Language Pairs
The speech encoder states are most important.
- Source text decoder helps to tune source speech encoder states.
- Source text helps find the right segment on source speech states.

Limitations
- Use a simple encoder-decoder model to translate input speech to target speech.
- It’s challenging to handle difficult translation tasks due to simple architecture.
- All modules use RNN.
Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder

- **ASR and MT decoders provide attention alignment necessary in the next block in the inference step.**
  - The model uses a combination of 3 attentions for alignment source speech to target speech.
- **Transcoder converts different hidden states** in speech translation.
  - The model converts ASR and MT hidden states and MT and TTS hidden states.
- We revamp the **RNN-based speech-to-speech translation model** with the **Transformer**.
Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder

Alignment information is most important for Speech Translation.
- This model provides attention alignment necessary in the next block.
- Transcoder converts different hidden states in speech translation.

Limitations
- The model has deep architecture, and the training is difficult.
- We need additional training to convert ASR hidden states into MT hidden states.
Proposed:
Training mechanism for Transcoder

The difference between ASR and MT.
✓ ASR models input acoustic phoneme sequences.
✓ MT models input word sequence.
✓ Each hidden states have different distributions.

Transcoder’s training
✓ Transcoder gets ASR context sequence and encodes them with considering context information.
✓ A pre-trained MT encoder provides the target sequence.
Proposed: Training mechanism for Transcoder

We extend the architecture and train model step by step.

✓ We get an idea from Curriculum Learning. We change the model architecture and task difficulty.
✓ During the training, the model becomes deep, and the task becomes difficult at each step.
✓ We utilize a pre-trained MT and TTS encoder as a teacher model.

**Step 1: Pre-train**
- ASR Encoder
- ASR Decoder
- MT Encoder
- MT Decoder
- TTS Encoder
- TTS Decoder

**Step 2: Train Transcoder**
- ASR Encoder
- ASR Decoder
- Transcoder
- English Speech
- Attention

**Step 3: Extend MT decoder**
- ASR Encoder
- ASR Decoder
- Transcoder
- English Speech
- Attention

**Step 4...**
- ASR Encoder
- ASR Decoder
- Transcoder
- English Speech
- Attention

- MT Decoder
- Konnichiawa

We extend the architecture and train model step by step.

- We get an idea from Curriculum Learning. We change the model architecture and task difficulty.
- During the training, the model becomes deep, and the task becomes difficult at each step.
- We utilize a pre-trained MT and TTS encoder as a teacher model.
**Experimental setups**

<table>
<thead>
<tr>
<th>Transformer setting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder layers</td>
<td>3</td>
</tr>
<tr>
<td>Decoder layers</td>
<td>6</td>
</tr>
<tr>
<td>Multi-head</td>
<td>8</td>
</tr>
<tr>
<td>Transformer hidden size</td>
<td>256</td>
</tr>
<tr>
<td>Transformer FFN [7] hidden size</td>
<td>1024</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RNN setting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder layers</td>
<td>4</td>
</tr>
<tr>
<td>Decoder layers</td>
<td>2</td>
</tr>
<tr>
<td>RNN type</td>
<td>LSTM</td>
</tr>
<tr>
<td>Multi-head</td>
<td>8</td>
</tr>
<tr>
<td>RNN hidden size</td>
<td>256</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Optimizer setting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Warm-up steps for Transformer</td>
<td>8000</td>
<td></td>
</tr>
<tr>
<td>Optimizer method</td>
<td>Adam</td>
<td></td>
</tr>
<tr>
<td>Learning rate decay</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Training steps</td>
<td>2,000,000</td>
<td></td>
</tr>
<tr>
<td>Batch size (ASR / MT / TTS / ST)</td>
<td>64 / 128 / 32 / 16</td>
<td></td>
</tr>
</tbody>
</table>

- ✓ Input is Japanese or English natural speech.
- ✓ Target is Japanese, English, Korean, or Spanish generated speech.
- ✓ We use the ASR to transcribe the generated Mel sequences and evaluate translation performance with BLEU and METEOR.
Experimental results

We evaluate ASR and MT performance to compare.

- RNN and Transformer model.
- Syntactically similar and distant language.

<table>
<thead>
<tr>
<th>Model</th>
<th>Natural English</th>
<th>Natural Japanese</th>
<th>Generated English</th>
<th>Generated Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>9.1</td>
<td>10.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Transformer</td>
<td>6.8</td>
<td>8.2</td>
<td>3.5</td>
<td>5.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Syntactic similar</th>
<th>Syntactic distant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En to ES</td>
<td>Ja to Ko</td>
</tr>
<tr>
<td>RNN</td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Transformer</td>
<td>45.9</td>
<td>62.1</td>
</tr>
</tbody>
</table>

- The Transformer model outperformed the RNN model on ASR and text MT tasks.
- The syntactic similar language translation achieved a higher score than syntactic distant language.
## Experiment Results

We evaluate Speech translation performance to compare.
- Proposed and Multi-task method.
- RNN and Transformer model.
- Syntactically similar and distant language.

### BLEU and METEOR scores of speech-to-speech translation

<table>
<thead>
<tr>
<th>Model</th>
<th>Syntactic similar</th>
<th></th>
<th>Syntactic distant</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>En to ES</td>
<td>Ja to Ko</td>
<td>En to Ja</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>METEOR</td>
<td>BLEU</td>
</tr>
<tr>
<td>Baseline: Cascade (RNN)</td>
<td>38.9</td>
<td>47.7</td>
<td>38.7</td>
<td>49.1</td>
</tr>
<tr>
<td>Baseline: Cascade (Transformer)</td>
<td>41.3</td>
<td>52.1</td>
<td>41.0</td>
<td>51.1</td>
</tr>
<tr>
<td>Multi-task (RNN)</td>
<td>38.8</td>
<td>48.2</td>
<td>39.1</td>
<td>49.9</td>
</tr>
<tr>
<td>Multi-task (Transformer)</td>
<td>43.1</td>
<td>58.8</td>
<td>42.5</td>
<td>58.3</td>
</tr>
<tr>
<td>Transcoder (Transformer)</td>
<td>44.0</td>
<td>59.3</td>
<td>42.9</td>
<td>58.8</td>
</tr>
</tbody>
</table>

- Our proposed approach outperforms Multi-task speech translation.
- The proposed method with Transformer further improved performances.
Analysis of the proposed and Multi-task methods

Both methods show monotonic attention alignments at ASR tasks. However, the Multi-task model's attention alignment is not clear.

Both MT attention alignments has a similar direction, as shown in the arrow.

The proposed method shows monotonic attention alignments again.

The Multi-task method shows week attention alignments similar to MT but not precisely the same.
Conclusion

• In this research, we propose the Transformer-based speech-to-speech translation with a transcoder that can pass the context information for each process.
  • Our proposed model improves BLEU and METEOR scores compared with the Multi-task model on syntactically distant language.

• Our proposed model learns the end-to-end speech translation step by step.
  • ASR and MT decoders support alignment at each step.

• Our proposed model shows the best performance on both syntactically similar and distant language pairs.
Thank you for your listening.